Chapter 1
The Era of Big Data: From Data-Driven Research to Data-Driven Clinical Care

Christian Baumgartner

Abstract When the era of big data arrived in the early nineteen nineties, biomedical research boosted new innovations, procedures and methods aiding in clinical care and patient management. This chapter provides an introduction to the basic concepts and strategies of data-driven biomedical research and application, an area that is explained using terms such as *computational biomedicine* or *clinical/medical bioinformatics*. After a brief motivation it starts with a survey on data sources and bioanalytic technologies for high-throughput data generation, a selection of experimental study designs and their applications, procedures and recommendations on how to handle data quality and privacy, followed by a discussion on basic data warehouse concepts utilized for life science data integration, data mining and knowledge discovery. Finally, five application examples are briefly delineated, emphasizing the benefit and power of computational methods and tools in this field. The author trusts that this chapter will encourage the reader to handle and interpret the huge amount of data usually generated in research projects or clinical routine to exploit mined bioinformation and medical knowledge for individualized health care.

Keywords Computational biomedicine • Data integration and management • Knowledge discovery • Data mining • Clinical applications

1.1 Introduction

In the past two decades, the new era of "big data" in experimental and clinical biomedicine has arrived and grown as a direct consequence of the availability of large reservoirs of data. Data collection in digital form was already underway by the 1960s, allowing for retrospective data management and analysis to be undertaken using computers for the first time. Relational databases arose in the 1980s along with Structured Query Languages (SQL), enabling dynamic, on-demand structural analysis and interpretation of data from complex research designs. The 1990s saw

C. Baumgartner, Ph.D. (✉)
Institute of Health Care Engineering with European Notified Body of Medical Devices, Graz University of Technology, Stremayrgasse 16, A-8010 Graz, Austria
e-mail: Christian.Baumgartner@TUGraz.at

X. Wang et al. (eds.), *Application of Clinical Bioinformatics*,
Translational Bioinformatics 11, DOI 10.1007/978-94-017-7543-4_1

an explosion in the growth of data associated with the emerging use of new high-throughput, lab and imaging technologies in fundamental biomedical research and clinical application. Data warehouses were beginning to be used for storing and integrating various types of data, where different data sources are transformed into a common format and converted to a common vocabulary needed to overcome computational challenges of data-driven research and development. The new era of "computational biomedicine" or "clinical bioinformatics" was born as a multidisciplinary approach that brought together medical, natural and computer sciences, aiming at uncovering unknown and unexpected biomedical knowledge stored in these data sources, which had the potential to transform our current clinical practices (Chang 2005; Wang and Liotta 2011; Coveney et al. 2014). Research areas such as data warehousing and information retrieval, machine learning, data mining, and others thus arose as a response to challenges faced by the computer science and bioinformatics community in dealing with huge amounts of data, enabling a better quality of data-driven decision making. As data are any facts, numbers, images or texts that can be accessed and processed by stand-alone computers or computational networks, the patterns, associations or relationships among available data can provide information about historical patterns and future trends so that undreamt of opportunities emerge for biomedical research and application. This knowledge may help to create a new way of dealing with clinical care and patient management never previously possible. Clinical bioinformatics, which resulted from the big data era, is thus a crucial element of the medical knowledge discovery process where relevant sources of medical information and bioinformation are combined and mined to allow for individualized healthcare.

1.2 The Revolution of High-Throughput and Imaging Technologies and the Flood of Generated Data

In the life sciences, huge amount of data are generated, utilizing the wide spectrum of high throughput and laboratory technologies, and modern health care imaging systems such as MRI or CT. In biomolecular research, microarray based expression profiling and more recently next-generation sequencing (NGS) technologies have become the methodology of choice e.g. for whole transcriptome expression profiling, producing a flood of data that need to be computationally processed and analysed (Worthey 2013; Soon et al. 2013). The most widely used NGS devices, for example, are able to sequence up to 150 bases from both sides of RNA fragments and create a maximum output of up to 1000 GB per run. Most advanced protein profiling technologies are implemented with a broad panel of mass spectrometry-based techniques to separate, characterize and quantify analytes from complex biological samples (Chen and Pramanik 2009; Brewis and Brennan 2010; Woods et al. 2014). Labs are typically equipped with diverse mass spectrometer (MS) systems including TOF-TOF, Quadrupole-TOF, FT-ICR, and LTQ-Orbitrap type analyzers. In this field, shotgun proteomics is a widely used tool for global analysis of protein

modifications, where, in a typical LC-MS/MS experiment, hundreds of thousands of tandem mass spectra are typically generated. Sophisticated computational tools for MS sprectra processing and database search strategies are used for the identification of peptide/protein modifications (Baumgartner et al. 2008; Cerqueira et al. 2010; Sjöström et al. 2015). In metabolomics, different fundamental approaches can be distinguished, i.e. untargeted and targeted metabolomics and metabolic fingerprinting (Baumgartner and Graber 2007; Putri et al. 2013; Naz et al. 2014; Zhang et al. 2015). Using targeted metabolomics, quantitiation of a preselected set of known metabolites by determining absolute values of analyte concentrations with the use of internal chemical standards allows for hypothesis-driven research and interpretation of data based on *a-priori* knowledge. To provide a holistic picture of metabolism, untargeted metabolic profiling aims at measuring as many analytes as possible (up to several hundreds) to create a snapshot of the biochemical profile within the analysed sample. The established technologies in metabolomics include – analoguous to proteomics – mass spectrometry based approaches and nuclear magnetic resonance (NMR) spectroscopy, generating thousands to tens of thousands data points per spectrum. Multiple processing steps are required to analyze this huge amount of spectral information, ranging from modalities for denoising, binning, aligning spectra to peak detection and high-level analysis e.g. for biomarker identification and verification (Swan et al. 2013; Netzer et al. 2015).

Nowadays bioimaging devices with increasing resolution are widely used in biological and clinical laboratories, generating imaging data with hundreds of Megabytes or Gigabytes (Eliceiri et al. 2012; Edelstein et al. 2014). Whole-slide bioimaging, for instance, combines light microscopy techniques with electronic scanning of slides and is able to collect quantitative data, currently regarded as one of the most promising avenues for diagnosis or prediction of cancer and other diseases. Traditional health care imaging technologies such as CT, MRI, ultrasound or SPECT and PET make it possible to assess the current status and condition of organs or tissues and to monitor patients over time for diagnostic evaluation or for controlling therapeutic interventions (Smith and Webb 2010; Mikla and Mikla 2013). In particular, CPU-intensive image reconstruction and modeling techniques allow instant processing of 2D signals to create 3D/4D image stacks of enormous amounts of data, typically stored in DICOM file format. This DICOM standard facilitates interoperability of medical imaging instrumentations, providing a standardized medical file format and directory structure, which enables access to the images and patient-related information for further processing, modeling and analysis.

1.3 Study Design and Data Privacy

Different epidemiological study designs such as case-control, (longitudinal) cohort studies or more complex designs such as randomized controlled trials are selected in biomedical research (Dawson and Trapp 2004; Porta 2014). Case-control studies,

which are always retrospective, are designed to determine if an exposure is associated or correlated with an outcome, i.e., a disease or biological/physiological condition of interest. This study type is referred to as an observational, non-experimental study where the investigator simply "observes", as the outcome of each subject enrolled in their respective groups is already known by the investigator. The investigator identifies the study groups of interest, i.e. cases (a group known to have the outcome such as patients with a coronary artery disease or patients with prostate cancer classified by the established gleason scoring scheme), and controls, which is a cohort known to be free of the outcome. Note that the same data must be collected in the same way from both groups. As the investigator usually makes use of previously collected data, a major limitation of observational studies is cofounding. By definition, a confounding variable is one which is associated with the exposure and is a cause of the outcome. For example, if researchers investigate whether "smoking" leads to "lung cancer", "smoking" is the independent variable and "lung cancer" is the dependent variable. Confounding variables are any other variables that also have an effect on the dependent variable like "age", which has a hidden effect on the selected dependent variables and may increase the variance in the data by introducing a bias. The big advantage and practical value of a case-control study is that this study design is very efficient, produces rapid results and may also be ideal for preliminary investigations of e.g. a suspected risk factor for a disease. However, results from case-control studies need to be independently verified and confirmed by larger, more accurate prospective cohort studies or randomized controlled clinical trials which are the only way to eliminate all confounding effects in a study. Randomized controlled trials, also known as double blind studies, are the most effective way of determining whether a cause-effect relation exists between a clinical intervention or treatment and outcome. Typically, subjects are allocated at random to receive one of several clinical interventions where one of them may serve as the standard or control investigation. This group usually receives a placebo or no clinical intervention. All intervention groups are treated identically. Although randomized controlled clinical trials are very powerful tools, they are time consuming and costly, and show some limitations by ethical and practical concerns such as recruitment and randomization. Besides the controlled collection of clinical information including biological samples, the acquisition of patient-related information is highly sensitive and data privacy is essential to prevent the unauthorized or unwanted disclosure of information about an individual if this data is not used for individual patient management and care. If this information is needed e.g. for biomedical research, educational purposes for the medical staff, etc., the strong data protection requirements can be overcome by approaches such as anonymization or pseudonymization (Neubauer and Riedl 2008; Elger et al. 2010). Anonymization, a method to disassociate all identifiers from the data, and pseudonymization, which supports an authorized re-identification of personalized data, make it possible to remove information from the data that are not strictly required for the intended purpose of those data and thus guarantees the privacy protection established by law.

1.4 Data Quality and Standard Operating Procedures (SOPs)

A Good Clinical Practice (GCP) is a central principle in biomedical research and clinical patient management, which allows for quality controlled collection and tracking of patient-related records, biosamples and additional study material. In the process of measuring and acquiring bioanalytical information gathered from biological samples such as blood (plasma, serum, dried spots), urine, other body fluids like sputum or lavages, cell cultures or tissue samples, the quality of generated data is crucial for the subsequent steps in data preprocessing and analysis. In general, the entire analytical and computational workflow, ranging from study design to study execution including pre-analytical sample handling, bioanalytical analyses, data aggregation and consolidation, computational tools for data integration, knowledge mining and interpretation, requires controlled procedures and standardized regulatory directives to ensure a high degree of consistency, completeness, and reproducibility of data and results. A guideline is provided by the "Guidance for Industry – Bioanalytical Method Validation" (FDA 2013) for the development of analytical methods and standard operating procedures (SOPs). According to laboratory-specific SOPs, a research lab has to handle a hazardous chemical safely and bring an application within the scope of the special execution procedures. These include the amount and concentration of proposed analytes, technical controls and inspections, as well as personal protective equipment with safety instructions. A controlled collection of samples, for example, is assured by the implementation of specific protocols and standards for sample taking using barcoding, a way to rapidly, accurately, and efficiently gather sample information and transmit it to a central data server for further analysis.

1.5 Life Science Data Warehouse Concepts for Data Integration and Knowledge Retrieval to Support Medical Decision Making

There is a strong need in life science research to integrate and store biomedical information generated in multiple research fields using proper approaches such as data warehouses (Parmanto et al. 2005; Töpel et al. 2008; Kienast and Baumgartner 2011; Hu et al. 2011; Lyne et al. 2013; Galhardas and Rahm 2014; Dander et al. 2014). By definition, data integration is the task of "combining the data residing at different sources, and providing the user with a unified view of the data" (Calì et al. 2001; 2003). These efforts require efficient and feasible IT concepts taking into account quality-assured standards and procedures (Shadbolt et al. 2006). The number, size, and complexity of life science databases continuously grow (Kei-Hoi et al. 2009) meaning that scientists in experimental and clinical research fields demand new concepts to handle (i) the variety and amount of available data,

(ii) data heterogeneity arising from different sources and (iii) a lack of standards for such integration concepts, which is a prominent problem (Kei-Hoi et al. 2009). Generally, heterogeneity in computer science can be divided into four classes, i.e. system heterogeneity (different hardware platforms and operation systems), semantic heterogeneity (differences in the interpretation of different data sources), syntactic heterogeneity (difference of data representation formats) and structural heterogeneity (different data models or structures). As a consequence of these subclasses of heterogeneity, the following challenges in life science data integration need to be taken into account (Kienast and Baumgartner 2011):

(1) The origin of data with different data formats. Basically, three groups of data structures can be defined:
   - structured data, which is organized in a form and structure so that it is identifiable. E.g. databases using Structured Query Language (SQL) for data management and retrieval.
   - semi-structured data which is used to identify certain elements within the data, but lacks a strict data model structure: E.g. metadata in HTML, XML formats and in other mark-up languages.
   - unstructured data that include other data types that are not part of a database: E.g. text (electronic patient records, biomedical literature), biomedical images of diverse formats like DICOM, JPG, TIF.

(2) The identification and interpretation of "synonyms" and "homonyms": In biomedicine scientists often name biological entities and relationships synonymously. For data integration it is crucial to strictly distinguish between synonyms, i.e. words that share meanings with other words, and homonyms i.e. words that sound similar, but have different meanings, and to select the right term in relation to the context. As examples, "entities" include anatomical terms like cells and tissues (cell and tissue types, anatomy, populations, etc.), biomolecules such as genes, proteins, metabolites (amino acids, enzymes, antibodies, protein genes, etc.) and "relationships" that include terms like gene expression, mutation, activation, inhibition, regulation, prognosis, diagnosis or therapy.
(3) The recognition of granularity: Biomedical data sources may provide information at different levels of granularity. For example one data source contains information about different metabolic diseases and their clinical phenotypes, symptoms and therapeutic recommendations while other databases provide detailed information about the same diseases, but characterized by the underlying molecular mechanisms, pathways and networks.
(4) The identification of viable resources: It is crucial to identify relevant and interoperable data sources that use widely accepted, comprehensive standards for the access and exchange of data to avoid unnecessary duplication and incompatibility in the collection, processing, and dissemination of such data.

In general, there are various approaches for integrating different data sources by using warehousing, mediation and Semantic Web technology based approaches

(Töpel et al. 2008; Pasquier 2008; Grethe et al. 2009; Spanos et al. 2012). Warehouse integration consists of cataloguing and accessing data from multiple sources and repositories in a local database, which is called the warehouse and designed with the objective of retrieving information from the data and supporting decision making (Fig. 1.1) (Töpel et al. 2008; Kienast and Baumgartner 2011). Usually relational databases with different database schemata are used such as the Star schema[1] or the Snowflake schema.[2] Basically, a data warehouse consists of two entities, a back room and front room entity, which are mostly separated physically as well as logically. The ETL process (Extract, Transform and Load), the basic concept of a data warehouse back room, is conducted to extract and import the data from the data source (i.e. data from external sources including flat files, XML files or databases) into the needed reference system (repository) of the data warehouse (Hernandez and Kambhampati 2004; Kugler et al. 2008). The back room thus describes the data management component, which permanently stores the data in a physical database and delivers subsets of data retrieved by queries. The front room makes it possible to access the data held in the warehouse by providing tools and methods for intelligent data accessing, mining and information retrieval, which is of great interest for clinical uses e.g. to support medical decision making. Users have access to the data of interest via the front room by two different types of database queries: ad-hoc and intelligent queries. Ad-hoc queries are executed through SQL statements that need to be formulated by a trained user, while intelligent query approaches are extended, more user-friendly queries based on result sets of adhoc queries using a medical knowledge base to process the information requests. Such requests allow for identification of patterns and relationships within the data relying on the concept of different information hierarchies in the data (Lyne et al. 2013).

Mediator based integration focuses more on query translation, where data is not centrally stored, but directly accessed from the distributed sources (Grethe et al. 2009). The data flow between mediators and data sources is provided by specific software components termed Wrappers (Hernandez and Kambhampati 2004). A more targeted approach for the integration of heterogeneous data sources, in particular in biomedical applications, is the use of Semantic Web technologies, whereby existing documents and data are provided with structured meta-information (Cheung et al. 2007; Pasquier 2008; Spanos et al. 2012). A key feature of this approach is the use of semantics by ontologies, which overcomes the problem of interpreting homonyms and synonyms in different sources. It should be noted that ontologies are a type of controlled vocabulary that attempt to capture the knowledge of a specific domain, which is also an important approach for warehousing and mediation based data integration concepts (Bodenreider 2008).

[1] It comprises so-called dimension tables containing data from different data sources and fact tables connecting various dimension tables.

[2] It uses a transformation of the dimension tables to the third normal form with less data space needed, but more complex data queries.

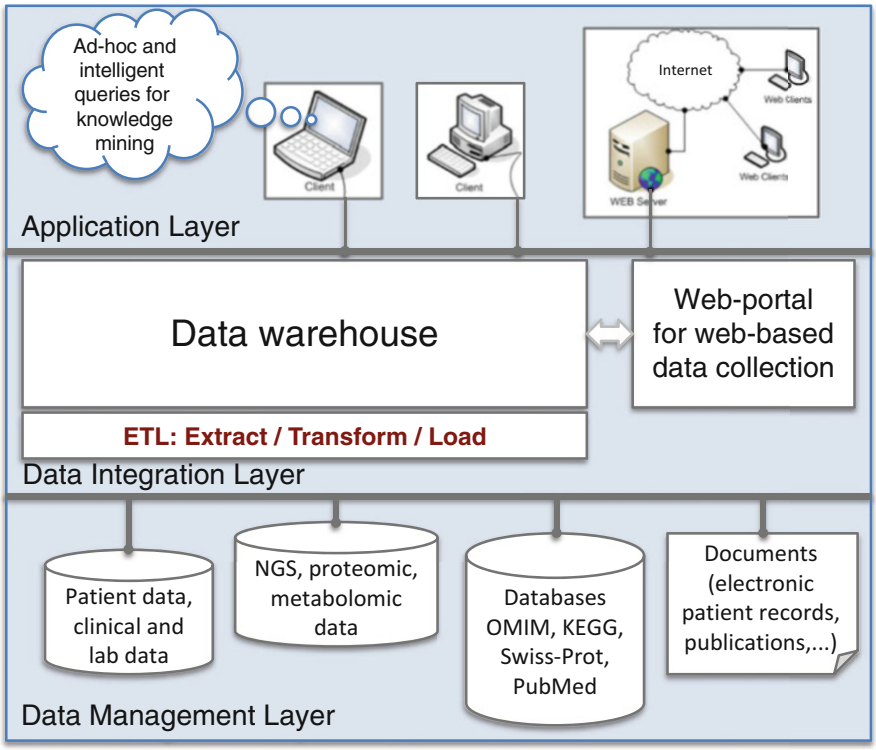

Fig. 1.1 Three-tier architecture of a data warehouse which is composed of a data management layer, data integration layer and application layer

To benefit from ontologies, it is important to annotate instances to metadata ontologies, which should be standardized and machine-readable. The Semantic Web technology provides tools for exchanging metadata information via Extensible Markup Language (XML) to allow for semantic data integration.

1.6 Knowledge Discovery and Data Mining for Clinical Care

According to Fayyad's 1996 definition (Fayyad et al. 1996a, b), knowledge discovery in general is the "nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". Data mining (DM) is a step in this process consisting of "particular data mining algorithms that, under some acceptable computationally efficiency limitations, produce a particular enumeration of patterns". The term knowledge discovery thus refers to the process of finding novel knowledge in the data. It does this by using data mining, machine learning or

biostatistical methods to extract and identify what is deemed knowledge, according to the specifications of measures and thresholds, using the given database with any required preprocessing, subsampling, and transformations of that data (Fig. 1.2). Note that the terms knowledge discovery and data mining are distinct. So this field is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems and data visualization, and requires an interdisciplinary view of research, in particular in a biomedical setting (Mitchell 1997; Pardalos et al. 2007; Ting et al. 2009; Dua and Chowriappa 2012; Holzinger and Jurisica 2014).

For the purpose of targeted analyses, data often need to be preprocessed and transformed into a standardized and quality assured format. Data preprocessing is used e.g. to normalize or rescale data (logarithmic scaling), to select data subsets, samples or single features or to remove outliers in the data in order to avoid manipulations of subsequently performed statistical analyses (Kotsiantis et al. 2006; García et al. 2015). Assuming a normal distribution of data, a common model for removing outliers is, for example, the use of the interquartile ranges. This simple statistical approach defines an outlier as observation outside the interquartile range.

Basically, data mining can be distinguished by the forms of supervised and unsupervised learning. In supervised learning or class prediction, knowledge of a particular class, group or study cohort is used to classify the class instances into the correct groups or to select significant features from the data in terms of high discriminatory ability or predictive value using a learning method. In biomarker discovery, the search for biomarker candidates is typically "supervised" because study groups in preclinical experimental studies or controlled clinical trials are typically well-defined and phenotyped. The data are then available in the form of tuples $T = \{(c_j, x) \mid c_j \mid C, x \mid X\}$, where c_j is the class label (e.g. normal, diseased, various stages of a disease, treated, etc.), and $X = \{x \mid x_1, \ldots, x_n\}$ is the set of given data (e.g. *metric* data such as lab measurements, gene expression data or mass spectral data, or *nominal* and *ordina*l data such as medical scores like the Glasgow Coma Scale). In this specific field, basic data mining and computational concepts for the search, prioritization and verification of biomarker candidates constitute filter-based feature selection algorithms or more sophisticated approaches such as embedded or ensemble methods (Lewis et al. 2008; Osl et al. 2008; Netzer et al. 2009; Millonig et al. 2010; Fang et al. 2012a, b; Swan et al. 2013, 2015;

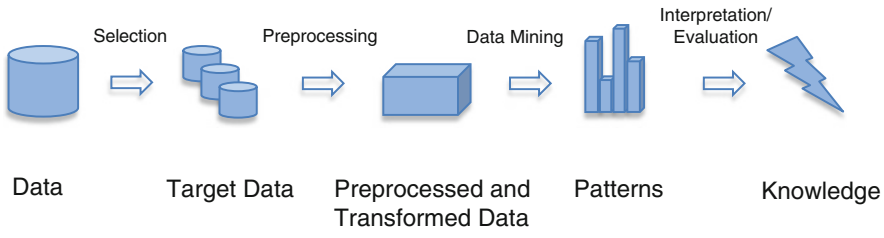

Fig. 1.2 General knowledge discovery process

Assawamakin et al. 2013). A survey of widely-used supervised feature selection techniques considering both independent and dependent samples w.r.t. unpaired and paired test hypotheses can be found in Baumgartner and co-authors (Baumgartner et al 2011). Newer approaches for identifying dynamic metabolic biomarkers using longitudinal or time-series data have been presented in Breit et al. (2015a, b).

The predictive performance and generalization power of validated clinical or biological markers is utilized to build classification models for medical decision making or disease screening. The basic idea of classification is to group or classify the given data $X = \{x \mid x_1, \ldots, x_n\}$ into the correct class $c_j \mid C$. For building classification models, multiple methods are available: logistic regression analysis, a widely-used method in biomedical applications, decision or classification trees, Bayes classifiers, k-nearest neighbor classifiers (k-NN), support vector machines, artificial neural networks or modalities of network-based approaches (Fielding 2006; Swan et al. 2013; Assawamakin et al. 2013). A high predictive value of such models is required to keep the false positive and false negative rate low which is expressed by high values of sensitivity and specificity (typically beyond 95-98 %). Note that such models have to consider the real incidence rate of a disease to correctly estimate the true false-positive rates. Statistical validation of a model is now the process of estimating how well a model, e.g. trained on a single derivation cohort, performs on future as-yet-unseen data by limiting problems like overfitting. Typical validation concepts include "train-and-test" strategies, splitting data into a separate train and test set, stratified cross-validation and permutation modalities, where the given data set is separated into train and test partitions. Multiple rounds of cross-validation are performed using the different partitions, and results are averaged over the rounds. This procedure reduces the variability in the data (Holzinger and Jurisica 2014).

In unsupervised data mining class information is unknown. A data set is typically given as a set of tuples in the form of $T = \{x \mid X\}$, where $X = \{x \mid x_1, \ldots, x_n\}$ is the set of given data (e.g. not annotated clinical scores, gene expression data, mass spectrometry data represented by lists of intensities vs. m/z values, voxels in a biomedical image, etc.). Using cluster analysis, data are grouped into meaningful classes based on similarity distance measures. Well-known methods are partitioning or hierarchical methods (k-means, single or average link) or newer methods that better consider local density structures in data such as DBSCAN (Ester et al. 1996) or Optics (Ankerst et al. 1999) as well as graph-based models (Fielding 2006; Xu and Wunsch 2010; Ye 2011). Association rule mining and regression analysis are complement methodological approaches in knowledge mining (Fig. 1.3).

To improve findings of single experiments, meta analysis as an further layer of integrated data analysis may be used, for example, for the search and verification of clinical and biological markers. This analysis strategy runs through multiple levels: (a) integrated analysis of the different clinical and preclinical experimental data which may arise e.g. from multi center studies and assessment of selected markers

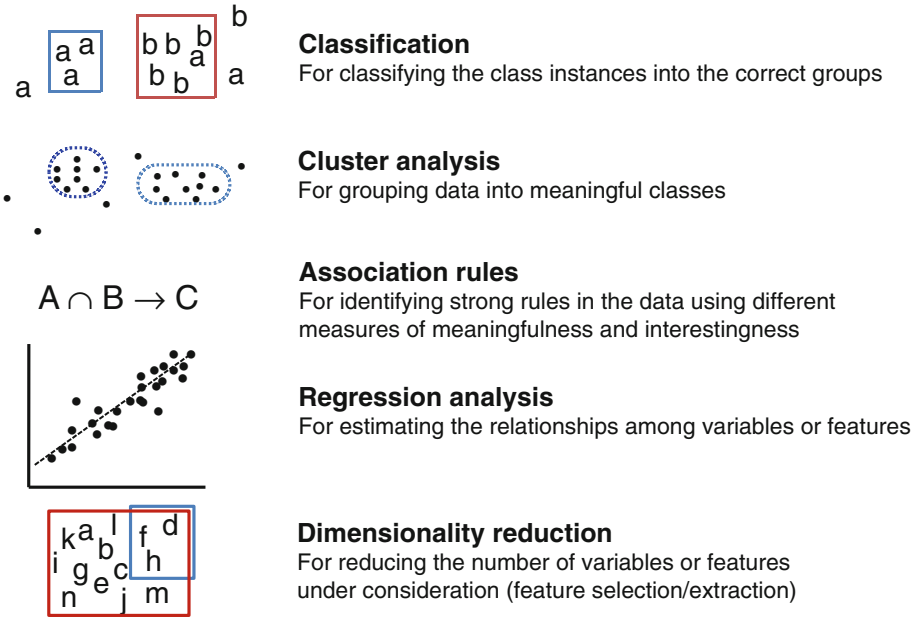

Fig. 1.3 Overview of the basic data mining areas used in biomedical research

with respect to their clinical plausibility and relevance, (b) application of statistical bioinformatics and data mining methods for searching and verifying marker candidates with superiour discriminatory ability with respect to the targeted patient cohorts or populations, (c) advanced correlation analysis by including all patient-relevant antropological, clinical and biomolecular data, (d) evaluation of predictive value of selected marker candidates by decision-outcome analysis and modeling. Tools such as Receiver operating characteristics (ROC) analysis and further approaches of health care technology assessment are used to estimate expected epidemiologic and economic consequences for individuals, the population and the public health (Mak et al. 2012; Tseng et al. 2012; Kaever et al. 2014).

1.7 Bio-Medical Application Examples

In the following section, selected application examples including studies from our research are presented, demonstrating the strength and benefit of computational approaches and concepts for data mining and knowledge discovery used in biomedical research and clinical care.

1.7.1 Example 1

Figure 1.4 shows an example workflow for the discovery of metabolic biomarkers in myocardical injury following the scheme of the general knowledge discovery process (Fayyad et al. 1996a, b, Fig. 1.2). In addition to a case/control study design, a longitudinal design for a biomarker search was selected where each subject serves as his/her own biological control. This design makes it possible to study the kinetic characteristics of circulating metabolites and thus to identify and classify biomarker candidates as early, late or sustained after acute injury (Lewis et al. 2008; Baumgartner et al. 2010).

1.7.2 Example 2

To support medical decision making in patients with Marfan syndrome (MFS), which is an autosomal dominant connective tissue disorder caused by mutations in the gene encoding fibrillin-1(FBN1) with highly variable clinical manifestations in the musculoskeletal, ocular and cardiovascular systems, a multiple logistic regression model was proposed by Baumgartner et al. (Baumgartner et al. 2005a, b). The model includes three cardiovascular parameters, i.e. the normalized diameters of aortic bulb and ascending aorta, and the ascending aortic distensibility. It demonstrated a sensitivity of almost 100 % and a specificity of 95 %, validated in an independent validation cohort. Interestingly, this model allows for the classification of patients with MFS only on three aortic parameters, selected from a pool of more than 30 measured parameters of the musculoskeletal, ocular and cardiovascular systems including genetic information (mutation data) (Baumgartner et al. 2005a, b).

Method Box
Classification model: A logistic regression model of the form $P = 1/(1+e^{-z})$ was selected, where p is the conditional probability $P(z=1 \mid x_1, ..., x_n)$ that MFS is present and z is the logit (discriminant function) of the model with three aortic parameters. A cut-off value ($P = 0.5$ by default) classifies controls if $P < 0.5$ and cases of disorder if $P \geq 0.5$. The logit of the MFS regression model is given by the following equation: $z = 4.379 + 2.293 \cdot$ normalized diastolic diameter of aortic bulbus [dimensionless] – $2.449 \cdot$ normalized diastolic diameter of ascending aorta [dimensionless] – $0.247 \cdot$ distensibility of ascending aorta [$kPa^{-1}\ 10^{-3}$] (see Baumgartner et al. 2005a, b).

For phenotype-genotype correlation, hierarchical cluster analysis on a collection of clinical symptoms of the MFS was performed. In this study, four phenotype classes (I, II, III, IV) could be identified, where the presence of missense

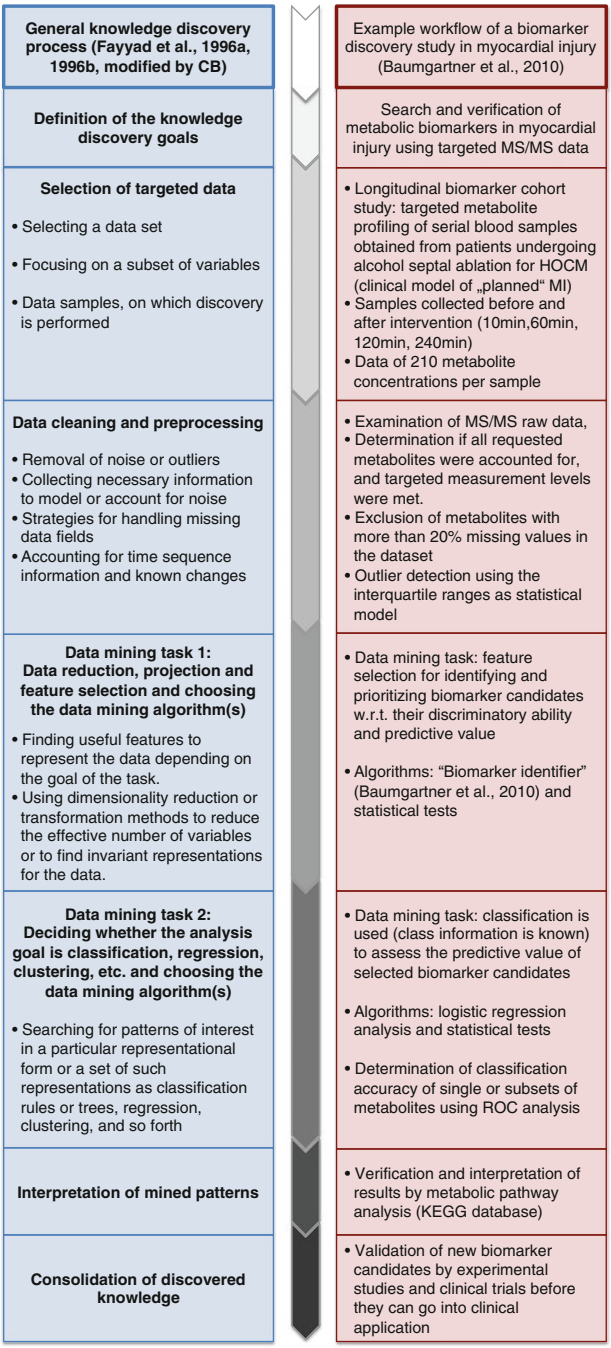

Fig. 1.4 Example workflow for the search for new biomarkers in myocardial injury based on the general scheme of the knowledge discovery process

mutations (substitutions or point mutations) was highly correlated with phenotype classes I and III, while phenotype class II, characterized e.g. by severe cardiovascular manifestations, was primarily associated with more complex mutations such as stop mutations or deletions with frame shift. These findings were used to specify characteristic clinical phenotypes with respect to different classes of mutations (substitutions vs. stronger forms of mutations such as stop mutations, insertions or deletions with frame shift) (see Fig. 1.5) (Baumgartner et al. 2005a, b, 2006).

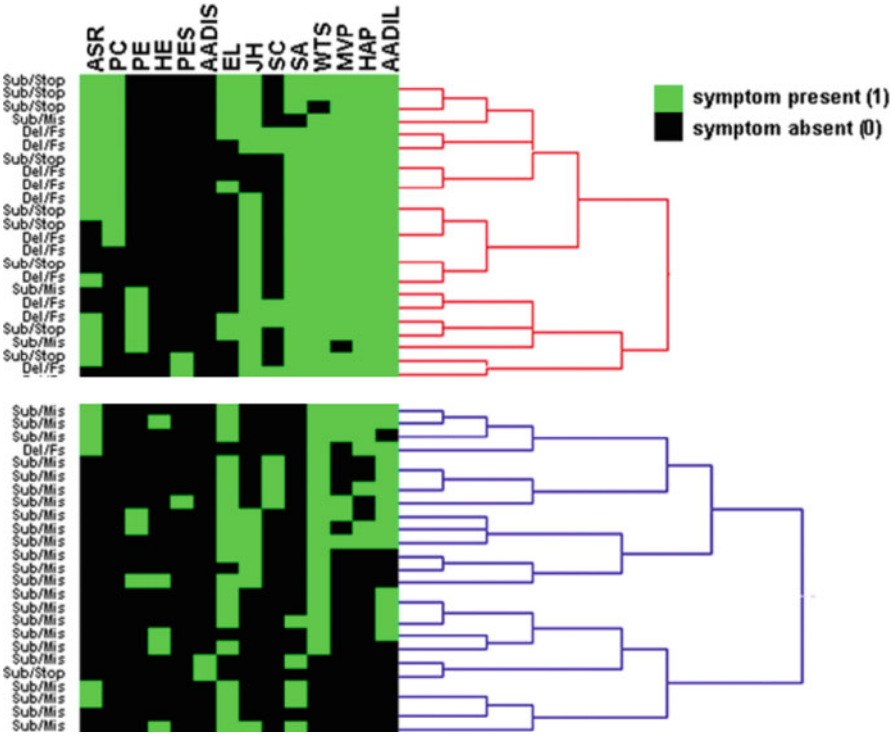

Fig. 1.5 Genotype-phynotype correlation in patients with MFS. Different classes of clinical phenotypes based on 14 examined clinical manifestations in the skeletal, ocular and cardiovascular system were identified using hierarchical cluster analysis. Each cluster was reviewed with respect to different forms of mutations (Sub/Mis . . . substitution-missense mutation, Sub/Stop . . . substitution with stop codon, Del/Fs . . . deletions with frameshift, Ins/Fs . . . insertions with frameshift) determined in each patient. Two different phenotype classes (I and II) with weak versus strong forms of mutations (substitutions (Sub/Mis) versus stronger forms of mutations such as stop mutations (Sub/Stop) or deletions with frame shift (Del/Fs) are shown in detail). Clinical manifestations: *EL* ectopia lentis, *AADIL* dilation of ascending aorta, *AADIS* dissection of ascending aorta, *MVP* mitral valve prolapse, *PC* pectus carinatum, *PES* pectus excavatum requiring surgery, *ASR* arm span ratio (ASR), *WTS* wrist or thumb sign, *SC* scoliosis, *PE* moderate pectus excavatum, *JE* joint hypermobility, *HAP* highly arched palate with crowding of teeth, *SA* striae atrophicae, *HE* herniae

Method Box
Cluster analysis: Hierarchical cluster analysis on 14 preselected clinical manifestations was performed to group patients with a similar clinical phenotype (clinical symptom present/absent, nominal measure). The average linkage method was selected, describing the distance between two clusters as the mean distance between an observation in one cluster and an observation in the other cluster. Clustered patient groups were reviewed with regard to their genetic predispositions (identified mutations). In the two-dimensional presentation of the clustered color map, each row represents a single mutation and each column the presence (green)/ absence (black) of examined clinical manifestations (see Fig. 1.5).

1.7.3 Example 3

In the work of Aronica and co-authors (Aronica et al. 2015) whole-genome expression profiles of 41 motor cortex samples of control and sporadic amyotrophic lateral sclerosis (SALS) patients were analyzed. ALS is a rapidly progressive neurodegenerative disease which is characterized by upper and lower motor neuron loss, leading to respiratory insufficiency and death after 3–5 years. Tissue samples were used for RNA extraction. A GenePix microarray scanner was selected to analyze gene expression changes in biological pathways associated with ALS. Although SALS patients could be clearly classified on the basis of their motor cortex gene expression profiles, no significant association between their clinical characteristics and cluster assignment was found.

Method Box
Cluster analysis: Hierarchical cluster analysis (similarity measure: Pearson centered; linkage method: average linkage, similarities were measured over the genes expressed in the motor cortex, 9646 genes in total) was used to separate and group controls from SALS patients. In the two-dimensional presentation of the clustered color map, each row represents a single gene and each column a motor cortex from controls or SALS patients. Colors indicate up-regulation, down-regulation or no change in gene expression.

1.7.4 Example 4

Wang and co-authors (Wang et al. 2014) introduced a comprehensive knowledge base, termed MitProNet, for the mitochondrial proteome, interactome and human disease associated mechanisms with mitochondria. This knowledge base allows for

a systematic identification of mitochondrial proteomes and a comprehensive characterization of functional linkages among mitochondrial proteins.

Method Box
Design of the knowledge discovery pipeline: A three step computational pipeline for data integration, modeling, analysis and interpretation was proposed. In step one, an inventory of mammalian mitochondrial proteins is integrated by collecting relevant proteomic datasets, and the proteins are classified using data mining and machine learning methods. A network of functional linkages among mitochondrial proteins is generated in step 2 by integrating 11 genomic features including protein-protein interaction, domain-domain interaction, shared domains, genomic context, genetic interaction, phenotypic semantic similarity, co-expression, GO semantic similarity, protein expression profiles, disease involvement and operon (operon contains a series of genes that are involved in the same biological process) based on the selected Naive bayes model. Step three prioritizes disease candidate genes by utilizing the network of functional linkages and network-based methods such as PageRank with Priors (PRP), Kstep Markov (KSM) or Heat Kernel Diffusion Ranking (HKDR). ROC analysis was selected for evaluating the performances of the various data sources and generated networks.

The system architecture and main contents of MitProNet can be found under doi:10.1371/journal.pone.0111187. The database is freely accessible.

1.7.5 Example 5

A web-based bioinformatics platform for clinical cancer research and routine applications in medical oncology, termed Personalized Oncology Suite (POS), integrating clinical data, NGS data and whole-slide bioimages from tissue sections was introduced by Dander and co-authors (Dander et al. 2014). Interestingly, POS combines biological data (mutations identified via next-generation sequencing and whole-slide bioimaging) and clinical data (information about the cancer patients, TNM staging, and density values of tumor-infiltrating lymphocytes used for immune score estimation) into one platform. As POS contains confidential and patient related data, the platform is secured by an authorization and authentication system (AAS). POS provides a convenient user interface, allowing for data upload, manipulation and visualization of integrated data. In a next release, POS will be extended by knowledge discovery and mining tools to aid in personalized cancer immunotherapy. The platform is open-source and can be downloaded at http://www.icbi.at/POS.

Method Box
Data warehouse application: The software architecture of POS as well as all detailed information on the functionality and provided features of the platform can be found under http://www.biomedcentral.com/content/pdf/1471-2105-15-306.pdf.

1.8 Is There a Need for Biomedical Scientists to Become Data Engineers?

Although not every biomedical scientist or clinician is a mathematician, statistician or computer scientist, it is definitely necessary to have basics skills in these fields. Modern study programs in medicine or biomedical sciences need to offer obligatory modules or tracks to strengthen these skills so that scientists are able to process, review and interpret data gathered in the scientist's field of expertise. Data-driven research approaches require all scientists in their field to develop and apply data analysis capabilities in computational biomedicine and statistics with proper sensitivity and quality assurance.

1.9 Conclusion

This chapter has provided a brief insight into concepts, methods, procedures, recommendations and applications in the field of computational biomedicine and was anticipated to appeal to those who undertake basic biomedical research or are interested in life science applications for clinical care. The presented sections were developed consecutively and cover the entire knowledge discovery process typically found in a biomedical setting. It might assist in the selection of computational methods and strategies for data collection, integration and analysis, which are urgently needed to transform mined knowledge into clinical applications.

References

Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering points to identify the clustering structure. In: Proceedings of the ACM SIGMOD international conference on management of data (SIGMOD'99), Philadelphia; 1999. p. 49–60.

Aronica E, Baas F, Iyer A, ten Asbroek AL, Morello G, Cavallaro S. Molecular classification of amyotrophic lateral sclerosis by unsupervised clustering of gene expression in motor cortex. Neurobiol Dis. 2015;74:359–76.

Assawamakin A, Prueksaaroon S, Kulawonganunchai S, Shaw PJ, Varavithya V, Ruangrajitpakorn T, Tongsima S. Biomarker selection and classification of "-omics" data using a two-step bayes classification framework. Biomed Res Int. 2013;2013:148014.

Baumgartner C, Graber A. Chapter 7: Data mining and knowledge discovery in metabolomics. In: Masseglia F, Poncelet P, Teisseire M, editors. Successes and new directions in data mining. Hershey, PA: Idea Group Inc; 2007. p. 141–66. ISBN 978-1-59904-639-6.

Baumgartner C, Mátyás G, Steinmann B, Baumgartner D. Marfan syndrome: a diagnostic challenge caused by phenotypic and genetic heterogeneity. Methods Inf Med. 2005a;44:487–97.

Baumgartner D, Baumgartner C, Mátyás G, Steinmann B, Löffler J, Schermer E, Schweigmann U, Baldissera I, Frischhut B, Hess J, Hammerer I. Diagnostic power of aortic elastic properties in young patients with Marfan syndrome. J Thorac Cardiovasc Surg. 2005b;129:730–9.

Baumgartner C, Mátyás G, Steinmann B, Eberle M, Stein JI, Baumgartner D. A bioinformatics framework for genotype-phenotype correlation in humans with Marfan syndrome caused by FBN1 gene mutations. J Biomed Inform. 2006;39:171–83.

Baumgartner C, Rejtar T, Kullolli M, Akella LM, Karger BL. SeMoP: a new computational strategy for the unrestricted search for modified peptides using LC-MS/MS data. J Proteome Res. 2008;7:4199–208.

Baumgartner C, Lewis GD, Netzer M, Pfeifer B, Gerszten RE. A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury. Bioinformatics. 2010;26(14):1745–51.

Baumgartner C, Osl M, Netzer M, Baumgartner D. Bioinformatic-driven search for metabolic biomarkers in disease. J Clin Bioinform. 2011;1:2.

Bodenreider O. Ontologies and data integration in biomedicine: success stories and challenging issues. In: Bairoch A, Cohen-Boulakia S, Froidevaux C, editors. Data Integration in the life sciences, Lecture notes in computer science, vol. 5109. Berlin/Heidelberg: Springer; 2008. p. 1–4.

Breit M, Baumgartner C, Weinberger KM. Chapter 9: data handling and analysis in metabolomics. In: Current applications of chemometrics. New York: Nova Sciences Publisher; 2015a. p. 181–203. ISBN: 978-1-63463-117-4.

Breit M, Netzer M, Weinberger KM, Baumgartner C. Modeling and classification of kinetic patterns of dynamic metabolic biomarkers in physical activity. PLoS Comput Biol. 2015b; 11(8): e1004454.

Brewis IA, Brennan P. Proteomics technologies for the global identification and quantification of proteins. Adv Protein Chem Struct Biol. 2010;80:1–44.

Calì A, Calvanese D, De Giacomo G, Lenzerini M. Accessing data integration systems through conceptual schemas, conceptual modeling – ER 2001, Lecture notes in computer science, vol. 2224. Berlin/Heidelberg: Springer; 2001. p. 270–84.

Calì A, Calvanese D, De Giacomo G, Lenzerini M. On the expressive power of data integration systems. In: Spaccapietra S, March S, Kambayashi Y, editors. Conceptual modeling – ER 2002, Lecture notes in computer science, vol. 2503. Berlin/Heidelberg: Springer; 2003. p. 338–50.

Cerqueira F, Graber A, Schwikowski B, Baumgartner C. MUDE: a new approach for optimizing sensitivity in the target-decoy search strategy for large-scale peptide/protein identification. J Proteome Res. 2010;9(5):2265–77.

Chang PL. Clinical bioinformatics. Chang Gung Med J. 2005;28(4):201–11.

Chen G, Pramanik BN. Application of LC/MS to proteomics studies: current status and future prospects. Drug Discov Today. 2009;14(9-10):465–71.

Cheung K, Smith A, Yip K, Baker C, Gerstein M. Semantic web approach to database integration in the life sciences. In: Baker CJO, Cheung K-H, editors. Semantic web. New York: Springer; 2007. p. 11–30.

Coveney P, Diaz V, Hunter P, Viceconti M. Computational biomedicine: modelling the human body. Oxford: Oxford University Press; 2014.

Dander A, Baldauf M, Sperk M, Pabinger S, Hiltpolt B, Trajanoski Z. Personalized oncology suite: integrating next-generation sequencing data and whole-slide bioimages. BMC Bioinf. 2014;15:306.

Dawson B, Trapp RG. Basic & clinical biostatistics (LANGE basic science). 4th ed. New York: Lange Medical Books/McGraw-Hill; 2004.

Dua S, Chowriappa P. Data mining for bioinformatics. Boca Raton: CRC Press; 2012.

Edelstein AD, Tsuchida MA, Amodaj N, Pinkard H, Vale RD, Stuurman N. Advanced methods of microscope control using μManager software. J Biol Methods. 2014;1(2):e10.

Elger BS, Iavindrasana J, Lo Iacono L, Müller H, Roduit N, Summers P, Wright J. Strategies for health data exchange for secondary, cross-institutional clinical research. Comput Methods Programs Biomed. 2010;99(3):230–51.

Eliceiri KW, Berthold MR, Goldberg IG, Ibáñez L, Manjunath BS, Martone ME, Murphy RF, Peng H, Plant AL, Roysam B, Stuurman N, Swedlow JR, Tomancak P, Carpenter AE. Biological imaging software tools. Nat Methods. 2012;9(7):697–710.

Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd International Conference. on Knowledge Discovery and Data Mining (KDD'96). Menlo Park: AAAI Press; 1996. p. 226–231.

Fang W, Chang X, Su X, Jian Xu, Zhang D, Ning K. A machine learning framework of functional biomarker discovery for different microbial communities based on metagenomic data. In: IEEE 6th International Conference on Systems Biology (ISB), Xiang, China, 2012a; p. 106–112.

Fang X, Netzer M, Baumgartner C, Bai C, Wang XD. Genetic network and gene set enrichment analysis to identify biomarkers related to cigarette smoking and lung cancer. Cancer Treat Rev. 2012b;2013(39):77–88.

Fayyad UM, Piatetsky-Shapiro G, Smyth P. Advances in knowledge discovery and data mining, chapter: from data mining to knowledge discovery: an overview. Menlo Park: AAAI Press. 1996a. p. 1–30.

Fayyad UM, Piatetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: towards a unifying framework. In: Simoudis E, Han JW, Fayyad UM (Hrsg.), editors. Proceedings of 2nd international conference on knowledge discovery and data mining, Portland, Oregon, AAAI Press; 1996b. p. 82–88.

FDA. Guidance for industry bioanalytical method validation. 2013. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm368107.pdf

Fielding AH. Cluster and classification techniques for the biosciences. Cambridge: Cambridge University Press; 2006.

Galhardas H, Rahm E. Data integration in the life sciences, Lecture notes in bioinformatics, vol. 8574. Berlin: Springer; 2014.

García S, Luengo J, Herrera F. Data preprocessing in data, mining, intelligent systems, Lecture notes in bioinformatics, vol. 72. Berlin: Springer; 2015.

Grethe JS, Ross E, Little D, Sanders B, Gupta A, Astakhov V. Mediator infrastructure for information integration and semantic data integration environment for biomedical research. Methods Mol Biol. 2009;569:33–53.

Hernandez T, Kambhampati S. Integration of biological sources: current systems and challenges ahead. SIGMOD Rec. 2004;33(3):51–60.

Holzinger A, Jurisica I. Interactive knowledge discovery and data mining in biomedical informatics, Lecture notes in computer science, vol. 8401. Berlin, Heidelberg: Springer; 2014.

Hu H, Correll M, Kvecher L, Osmond M, Clark J, Bekhash A, Schwab G, Gao D, Gao J, Kubatin V, Shriver CD, Hooke JA, Maxwell LG, Kovatich AJ, Sheldon JG, Liebman MN, Mural RJ. DW4TR: a data warehouse for translational research. J Biomed Inform. 2011;44 (6):1004–19.

Kaever A, Landesfeind M, Feussner K, Morgenstern B, Feussner I, Meinicke P. Meta-analysis of pathway enrichment: combining independent and dependent omics data sets. PLoS ONE. 2014;9(2):e89297.

Kei-Hoi C, Robert F, Scott M, Matthias S, Jun Z, Adrian P. A journey to semantic web query federation in the life sciences. BMC Bioinf. 2009;10 Suppl 10:S10.

Kienast R, Baumgartner C. Chapter 3: data integration on biomedical data using semantic web technologies In: Mahdavi MA, editors. Bioinformatics/Book 1, ISBN 978-953-307-282-1. Rijeka: InTech Open Access Publisher; 2011; p. 57–82.

Kotsiantis SB, Kanellopoulos D, Pintelas PE. Data preprocessing for supervised leaning. Int J Elect Comput Eng. 2006;1:2.

Kugler K, Tejada M, Baumgartner C, Tilg B, Graber A, Pfeifer B. Bridging data management and knowledge discovery in the life sciences. Open Bioinform J. 2008;2:28–36.

Lewis GD, Wei R, Liu E, Yang E, Shi X, Martinovic M, Farrell L, Asnani A, Cyrille M, Ramanathan A, Shaham O, Berriz G, Lowry PA, Palacios I, Tasan M, Roth FP, Min J, Baumgartner C, Keshishian H, Addona T, Mootha VK, Rosenzweig A, Carr SA, Fifer MA, Sabatine MS, Gerszten RE. Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury. J Clin Invest. 2008;118:3503–12.

Lyne M, Smith RN, Lyne R, Aleksic J, Hu F, Kalderimis A, Stepan R, Micklem G. metabolicMine: an integrated genomics, genetics and proteomics data warehouse for common metabolic disease research. Database (Oxford). 2013;2013:bat060.

Mak A, Cheung MW, Fu EH, Ho RC. Meta-analysis in medicine: an introduction. Int J Rheum Dis. 2010;13(2):101–4.

Mikla VI, Mikla VV. Medical Imaging Technology, 1st ed. Waltham: Elsevier; 2013. ISBN: 9780124170216.

Millonig G, Praun S, Netzer M, Baumgartner C, Mueller S, Villinger J, Vogel W. Non-invasive diagnosis of liver diseases by breath analysis using an optimized ion-molecule reaction-mass spectrometry approach: a pilot study. Biomarkers. 2010;15(4):297–306.

Mitchell TM. Machine learning. Boston: McGraw-Hill; 1997.

Naz S, Vallejo M, García A, Barbas C. Method validation strategies involved in non-targeted metabolomics. J Chromatogr A. 2014;1353:99–105.

Netzer M, Millonig G, Osl M, Pfeifer B, Praun S, Villinger J, Vogel W, Baumgartner C. A new ensemble based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry. Bioinformatics. 2009;25(7):941–7.

Netzer M, Handler M, Pfeifer B, Dander A, Baumgartner C. Knowledge discovery in proteomic mass spectrometry data. In: Tran QN, Arabnia H, editors. Emerging trends in computational biology, bioinformatics, and systems biology – algorithms and software tools. Waltham: Elsevier/MK; 2015. ISBN 9780128025086.

Neubauer T, Riedl B. Improving patients privacy with pseudonymization. In: Proceedings of the international congress of the European Federation for medical informatics. Studies in health technology and informatics, vol 136. Amsterdam: IOS Press; 2008. ISBN: 978-1-58603-864-9.

Osl M, Dreiseitl S, Pfeifer B, Weinberger K, Klocker H, Bartsch G, Schäfer G, Tilg B, Graber A, Baumgartner C. A new rule-based data mining algorithm for identifying metabolic markers in prostate cancer using tandem mass spectrometry. Bioinformatics. 2008;24:2908–14.

Pardalos PM, Boginski VL, Vazacopoulos A. Data mining in biomedicine. Berlin: Springer; 2007.

Parmanto B, Scotch M, Ahmad S. A framework for designing a healthcare outcome data warehouse. Perspect Health Inf Manag, 2005;2:3.

Pasquier C. Biological data integration using semantic web technologies. Biochimie. 2008;90 (4):584–94.

Porta M. A dictionary of epidemiology. 5th ed. Oxford: Oxford University Press; 2014.

Putri SP, Yamamoto S, Tsugawa H, Fukusaki E. Current metabolomics: technological advances. J Biosci Bioeng. 2013;116(1):9–16.

Shadbolt N, Hall W, Berners-Lee T. The semantic web revisited. IEEE Intell Syst App. 2006;21 (3):96–101.

Sjöström M, Ossola R, Breslin T, Rinner O, Malmström L, Schmidt A, Aebersold R, Malmström J, Niméus E. A combined shotgun and targeted mass spectrometry strategy for breast cancer biomarker discovery. J Proteome Res. 2015;14(7):2807–18.

Smith BS, Webb A. Introduction to medical imaging: physics, engineering and clinical applications (Cambridge texts in biomedical engineering). Cambridge: Cambridge University Press; 2010. ISBN 978-0521190657.

Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. Mol Syst Biol. 2013;9:640.
Spanos DE, Stavrou P, Mitrou N. Bringing relational databases into the semantic web: a survey. J Sem Web. 2012;3(2):169–209.
Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. OMICS. 2013;17(12):595–610.
Swan AL, Stekel DJ, Hodgman C, Allaway D, Alqahtani MH, Mobasheri A, Bacardit J. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. BMC Genomics. 2015;16(Suppl 1):S2.
Ting SL, Shum CC, Kwok SK, Tsang AHC, Lee WB. Data mining in biomedicine: current applications and further directions for research. J Softw Eng Appl. 2009;2:150–9.
Töpel T, Kormeier B, Klassen A, Hofestädt R. BioDWH: a data warehouse kit for life science data integration. J Integr Bioinform. 2008;5(2):93.
Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012;40(9):3785–99.
Wang XD, Liotta L. Clinical bioinformatics: a new emerging science. J Clin Bioinform. 2011;1 (1):1.
Wang J, Yang J, Mao S, Chai X, Hu Y, et al. MitProNet: a knowledgebase and analysis platform of proteome, interactome and diseases for mammalian mitochondria. PLoS ONE. 2014;9(10): e111187.
Woods AG, Sokolowska I, Ngounou Wetie AG, Wormwood K, Aslebagh R, Patel S, Darie CC. Mass spectrometry for proteomics-based investigation. Adv Exp Med Biol. 2014;806:1–32.
Worthey EA. Analysis and annotation of whole-genome or whole-exome sequencing-derived variants for clinical diagnosis. Curr Protoc Hum Genet. 2013;79:Unit 9.24.
Xu R, Wunsch 2nd DC. Clustering algorithms in biomedical research: a review. IEEE Rev Biomed Eng. 2010;3:120–54.
Ye Xiao-rong. Analysis on network clustering algorithm of data mining methods based on rough set theory. 2011 fourth international symposium on Knowledge Acquisition and Modeling (KAM), Sanya, 8–9 October. 2011; p. 296–298. ISBN: 978-1-4577-1788-8.
Zhang A, Sun H, Yan G, Wang P, Wang X. Mass spectrometry-based metabolomics: applications to biomarker and metabolic pathway research. Biomed Chromatogr. 2016;30(1):7-12.

Christian Baumgartner, PhD is Professor of Health Care Engineering at Graz University of Technology, Austria. He received his MSc (1994) and PhD degree (1998) in Electrical and Biomedical Engineering from Graz University of Technology, Austria, and his habilitation degree (Assoc.-Prof.) in Biomedical Engineering from the University for Health Sciences, Medical Informatics and Technology (UMIT), Hall in Tyrol, Austria (2006).

From 1998 to 2002, Dr. Baumgartner held an R&D position at Tecan.com, where he developed confocal fluorescence laser scanning systems for micro array applications. From 2007–2008, he was part of the Barnett Institute of Chemical and Biological Analysis, Northeastern University and Harvard Medical School, Boston, MA, where Dr. Baumgartner worked in the field of computational biomarker discovery. In 2009, he was appointed Full Professor, Director of the Institute of Electrical and Biomedical Engineering, and Vice Chair of the Department of Biomedical Informatics and Mechatronics at UMIT. He has been a professor since 2015 and Head of the Institute of Health Care Engineering with European Notified Body of Medical Devices at Graz University of Technology in Austria since 2016.

Dr. Baumgartner is the author of more than 150 publications in refereed journals, books and conference proceedings, and is a reviewer for more than 35 scientific journals. He served as a deputy editor of the “Journal of Clinical Bioinformatics”, and is an editorial board member of “Clinical and Translational Medicine”, “Methods of Information in Medicine” and “Cell Biology and Toxicology”. His main research interests include cellular electrophysiology, biomedical sensors and signal processing, biomedical modeling, simulation, clinical bioinformatics and computational biology.